Journal of Experimental Criminology
https://doi.org/10.1007/s11292-019-09399-6

The trick does not work if you have already seen the gorilla: how anticipatory effects contaminate pre-treatment measures in field experiments

Barak Ariel[1,2] · Alex Sutherland[3] · Matthew Bland[1]

Published online: 27 December 2019

Abstract

Objectives If participants can anticipate the intervention, they may alter their responses prior to exposure to treatment. One often-ignored consequence of these "anticipatory effects" (AE) is an impact on the pre-treatment measurement. We explore this potential contamination and present practical options for mitigating AE.
Methods A multidisciplinary review of AE.
Results Pre-treatment measures, especially pre-treatment dependent variables, can be contaminated by AE. Experimenters need to understand the following: (1) When did the treatment 'commence'? (2) How is the pretest measured? (3) Are AE specific or global? (4) What conclusions can we draw where pretest measures are contaminated by AE?
Conclusions AE are often ignored for both research and policy, which may lead to erroneous conclusions regarding effectiveness, benefits being underestimated, or both. AE can be resolved by collecting 'clean' baseline measures prior to the commencement of the AE, but the first step is to be aware of the potential bias due to this treatment × pre-measurement interaction.

Keywords Anticipatory effects · Pretest measures · Bias · Experiments

Introduction

In causal research, scholars often collect pretest scores as a measure of the participants' responses prior to the administration of the intervention. This 'pretesting' is used to

✉ Barak Ariel
ba285@cam.ac.uk

[1] Institute of Criminology, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, UK

[2] Institute of Criminology, Faculty of Law, Hebrew University, Mount Scopus, 91905 Jerusalem, Israel

[3] Behavioural Insights Team, 4 Matthew Parker St, Westminster, London SW1H 9NP1, UK

establish a baseline before the stimulus is introduced (Entwisle 1961; Richland et al. 2009). The pretest can also be used to check on the equivalence of the groups (however, cf. Gruijters 2016; Harvey 2018; Moher et al. 2012); to stratify participants on the basis of the pretest scores; to provide covariates in quasi-experimental designs (Kim and Willson 2010: p. 744); and to examine the nature of attrition in the follow-up period (Haverman 1987). Thus, 'pretreatment information, preferably on the same dependent variable used at posttest, helps enormously in answering such questions' (Shadish et al. 2002: p. 260).

Importantly, true population means of the 'pre-treatment conditions' are, in principle, meant to be unconditioned by the treatment effect. The underlying premise is that the pretest measures characterise the test and control subjects under no-treatment conditions, with the assumption that the scores define how the participants behave under 'natural' settings before the experiment (Farrington et al. 2003: p. 16–17; Gribbons and Herman 1997; Sprangers and Hoogstraten 1989). The treatment effect is subsequently presumed to be independent of these pre-treatment measures, and vice versa. True baseline mean scores are not meant to reflect responses to the treatment, and the within-subjects measures are indicative of a baseline measure that predates any exposure to the tested manipulation (Shadish et al. 2002: p. 260).

However, when measures do not follow the temporal sequence of pretest-intervention-posttest, then the causal inference is said to be confounded, and changes between the two data points (i.e. before and after) do not depict the true population variations caused by the intervention relative to control conditions (Campbell et al. 1963; Rossi et al. 1980). Our main purpose here is to illustrate how experimental designs that utilise pretest measures may miss a critical interaction effect between the baseline measure and the treatment effect. Anticipating the treatment, the survey questions, the rationale of the test or the contextual factors that surround the experiment can all influence the effect of the stimulus. When participants experience enhanced awareness in response to a future stimulus, they exhibit variations in skills, ability or behavioural manifestation, as if the stimulus is already present prior to the pretest measure. These 'anticipatory effects' introduce systematic noise because the pretest population mean scores are under- or overestimated. Moreover, any measure of the before-after causal estimate is inherently biased and can produce inaccurate effect sizes or an inappropriate conclusion about the null hypothesis, and ultimately miss out on beneficial interventions which may, in fact, be useful.

Researchers in different disciplines have unearthed a wide range of manifestations of anticipatory effects. We have evidence from crime control studies (Smith et al. 2002; Gibbs 2010), social psychology (Edwards 1957) and visual memory research (Hanslmayr et al. 2009). In crime prevention studies, for example, the effect appears to be quite pronounced, when 'publicity about new measures and preparations for their introduction, results in decreases of crime immediately before they are actually implemented' (Perry et al. 2017: p. 732), or 'in response to information about changes in enforcement presence and activities' (Gibbs 2010: p. 3). Furthermore, most textbooks on research methods cover anticipatory effects at least to some degree as 'response distortion', namely social desirability or halo effects (Edwards 1957), whereby participants react to manipulations as a function of what they perceive to be the 'expected' response rather than their usual perceptions, behaviour or propensity. There is also a robust body of literature that shows how test scores are influenced by exposure to earlier versions of the test (Crowne and Marlowe 1960). Thus, anticipatory effects, even if identified by different terms, are ubiquitous and well known.

However, a review of the influential randomised trials meta-analysed in our field (see, for example, Weisburd et al. 2016) demonstrates clearly that experimenters rarely consider the degree to which pretest sample measures reflect no-treatment population means at baseline. Experiments on the effect of various interventions, including policing, corrections, school-based initiatives and crime reporting, for example, frequently make an assumption about the unbiased nature of the baseline measurement which, in both theory and principle, may be unwarranted in field experiments. For example, premature exposure to the pretest may sensitise participants in unforeseen ways, and their performance on the posttest measure may be due to confounding variations in the pretest rather than to the treatment.

In this research note, we highlight these anticipatory effects in experiments, as we believe scholars do not appreciate just how problematic they can be. We review the relevant literature that addresses these concerns and offer a crude yet useful solution to the matter.

Key questions for researchers

Issues with anticipatory effects and the lack of problematising of pretesting biases in experimental studies highlight at least four key concerns, which we formulate as four empirical questions that experimenters should consider.

When can we say that the treatment effect 'commenced'?

Under the customary experimental model, all randomised units in the groups are observed before and after that moment of random allocation. Whether scholars are using the intention-to-treatment model—where participants are analysed based precisely on the random allocation sequence, regardless of their compliance with entry criteria and protocol-treatment actually received or experimental morality (Hollis and Campbell 1999; see earlier versions in Peto et al. 1977), or any other alternative design (Sheiner and Rubin 1995), the assumption is that the treatment commences from the instant of random allocation into treatment and control conditions.

However, there is no fundamental theoretical justification for this approach when we consider what is known about anticipatory effects. Indeed, having a single cut-off time from which the stimulus begins is elegant, convenient and robust, at least from a statistical perspective. Nonetheless, assuming participants do not react and experience variations in their opinions, emotions or behaviours, *before* the stimulus appears, is usually unwarranted, at least in field trials. Rarely do we see experiments in criminology in which the programme of change is 'parachuted', where the key players—the treatment providers, police officers, judges, offenders or members of the public—are suddenly participants in a controlled study. The treatment commences earlier than the physical administration of the 'intervention' (i.e. at the moment of the random allocation sequence) and is much more likely to have an effect at the moment of first awareness of the anticipated arrival of the intervention.

The 'start' date can be when participants are invited/told to participate in an intervention. It can also be when they are made aware that the specific treatment is possible for them, or when they become cognisant that individuals who share similar attributes to them have received, or are potentially receiving, the treatment of interest.

Although we acknowledge that the physical administration of the intervention carries the most weight, we must also appreciate the magnitude of the anticipation effect, which should not be overlooked. We argue that the 'treatment' commences with the anticipation of the treatment, rather than the sensual exposure to the stimulus.

What forms the pretesting measure?

Directly linked to the temporal segmentation we identified regarding the definition of 'treatment', we can also ask what the pretest measure represents. Again, most statistical models for the pretest-posttest experimental designs assume a single instant in time (i.e. the moment of measurement) unaffected by the stimulus. However, if the proposition that participants indeed anticipate the intervention and alter their perceptions and behaviours ahead of time is true, then pretesting can be confounded with the treatment effect as well. The 'baseline' measure can thus be complex and contaminated when participants expect to be involved in an intervention/experiment and only *then* are measured at pretest.

Consequently, experimenters need to consider whether the baseline measures were collected sufficiently early or whether they mask an interaction effect with the treatment variable. At the same time, the baseline measures need to be congruent with the experimental programme, so the measure cannot be outdated in relation to the stimulus. For example, if historical changes occurred in law, training, costumes or recording practices, then the pretest measure may be irrelevant if it reflects archaic periods.

That said, under certain conditions, it may not be possible to measure 'pure' baseline levels. For example when conducting experiments in police training (e.g. Antrobus et al. 2019), or collecting pretest measures in correction settings (e.g. Franke et al. 2010) or public perceptions regarding new policy initiatives (e.g. Hohl et al. 2010), the pretest measure is often collected immediately before the administration of the treatment condition, if measured at all. In these circumstances, we recommend that at the very least, scholars acknowledge this problem if quantification of the anticipatory effect is not possible.

Are anticipatory effects case-specific or are there industry-wide, global anticipatory effects?

There are two levels to this problem: case-specific, which is related to one or a series of related tests, or global, which refers to anticipatory effects that carry over from one remote study to other locales.

Case-specific anticipatory effects Single-site anticipatory effects are not a new phenomenon to criminological researchers (see Smith et al. 2002; see more broadly in Campbell et al. 1963: pp. 13–25; Kent et al. 1974; Michelson et al. 1985). Equally, there is the chance of anticipatory backfire, as described by Linning et al. (2019) in their discussion of the wider framework of temporal effects of interventions. We argue that in all cases, the 'package' or the mechanism by which the treatment is supposed to 'work' is a key variable to be understood in the interpretation of treatment effects.

More substantively, in the case-specific contamination of the pretest measure, participants react to the intervention *before* the pretest measure is taken, in anticipation of their own exposure to the stimulus. For example, offenders predict a future increased police intervention in the hot spots where they operate even before the police patrol the

area and even before the baseline measure of crime in the hot spots is collected and adjust their behaviour accordingly (see Ariel and Partridge 2017). Awareness of the anticipated intervention may come from the media, experience or a police ritual (see Gibbs 2010; Goldkamp and Vilcica 2008; Johnson and Bowers 2003). For instance, Boba and Roberto Santos (2007) stipulated that a programme of change in Florida to reduce construction site theft experienced reduced theft during the planning stage before the project was fully implemented because thieves adjusted their risk perceptions. Therefore, offenders stop or delay criminal activity as the specific target area has become too risky (Weisburd et al. 2006).

Global anticipatory effects We contend that a neglected consequence of anticipatory effects is their general reach. In part, we can contextualise these global anticipatory effects within the reproducibility and replicability concern in the social sciences and our field more specifically (Aarts et al. 2015; Barnes et al. 2019; Farrington et al. 2018, b; McNutt 2014; Pashler and Wagenmakers 2012; Pridemore et al. 2018; Świątkowski and Dompnier 2017). We argue that, because we live in the internet age, rapid exposure to groundbreaking research is common (see, for example, Allen et al. 2013; Larivière et al. 2015; Osca-Lluch and Haba 2005, see also Lehrer et al. 2007), just as it has been but with a more limited capacity with the mass media for some time now (Stocking and Dunwoody 1982). Therefore, there is a risk that pretest measures are affected by these publications.

Science needs to be disseminated to the public. However, the dissemination has a potential effect on the baseline scores in at least two ways. First, any experiment that includes a degree of deception (i.e. surprise) will likely fail replication if the participant is already cognizant of the unexpected manipulation. Take, for example, one of the most celebrated recent studies in psychology on the disappearing gorilla experiment (Simons and Chabris 1999). Participants were shown a video and tasked with counting how many times three basketball players wearing white t-shirts passed a ball. Approximately 30 seconds into the video, a person in a gorilla suit walked into the scene, faced the camera, thumped his/her chest and walked off the screen, spending a total of 9 seconds on screen. Half the viewers missed the gorilla (Simons and Chabris 1999). The study illustrated 'inattentional blindness'—people's inability to detect unexpected objects to which they are not paying attention—and highlighted major concerns about eyewitness testimony and the potential for wrongful convictions (Brewer and Palmer 2010). However, replications of this experiment failed to produce a similar result, for an obvious reason: the video was an Internet sensation, and once viewers expected the gorilla to make an appearance, most viewers did not miss it.

The lesson for our purposes should also be clear: when the anticipated intervention is novel and potent, then *other* future participants who can potentially be exposed to the stimulus may change their behaviour even before the intervention is introduced to them. To demonstrate this 'global anticipatory effect', consider its impact on research on police body-worn cameras (BWCs). BWCs are mobile, on-person cameras that police officers attach to their uniforms to record face-to-face interactions with members of the public. Beyond their ability to collect incriminating evidence, the key objective of this technology is to produce a 'civilising' effect: people behave differently when they are being recorded, and awareness of a 'watching eye' is hypothesised to reduce police violence, as officers and members of the public do not want to get caught abusing their powers and applying

unnecessary force to suspects. Importantly, BWCs have attracted public attention following the Rialto CA experiment (Ariel et al. 2015, 2019); thus, they are considered as a strong potential apparatus in assisting police-public relations (President's Task Force on 21st Century Policing 2015).

The message to police everywhere was clear: 'wear BWCs to stop you from misusing your powers' (Graham et al. 2019; Lee et al. 2019; Obama 2016: pp. 864–865). As a result, many police departments have, over time, equipped most of their frontline staff with BWCs. Public and professional interest in BWCs has mirrored a steady growth in scientific inquiry regarding their effects on police use of force and complaints made by members of the public against police officers (Lum et al. 2019). The major issue, however, is that initially, a number of early studies across different geographies and cultures revealed significant and often substantial reductions in police use of force following the introduction of BWCs (Ariel et al. 2015; Ellis et al. 2015; Goodall 2007; Jennings et al. 2015; Mesa Department 2013). However, later studies have failed to reproduce these effects, detecting nonsignificant results in terms of the same measure of use of police force (Braga et al. 2018; Peterson et al. 2018; White et al. 2017; Yokum et al. 2017).

We believe that the global anticipatory effect can partly explain the lack of congruence between early and later studies in a major way: officers adjusted their behaviour even before they were issued BWCs; thus, the studies have suffered from a measurement-interaction effect (see implications in Farrington 1983, as well as McCambridge et al. 2011). We think that officers (and possibly members of the public too) changed their demeanour *prior* to the pretest measure. In later tests of BWCs, the pretest scores were lower than they 'should' be, as participants adjusted their behaviour before the pretest measure was taken. The magnitude of the change, over time, was lower than expected when the pretest measure is unbiased. which manifested itself in lower counts of complaints, use of force, arrests, etc., in the pretest measure. In other words, although early studies in which the treatment (i.e. BWCs) was a 'surprise' detected significant pretest-posttest effects on the outcome measure, the pretest in later studies was contaminated because the change in behaviour already existed prior to the baseline measure. They have already seen the gorilla.

What can we conclude from studies whose pretest measures are affected by the anticipated treatment effect?

What determines our judgement about the usefulness of the pretest measure is whether it meets our purposes. Scholars wish to be sufficiently protected from making false claims about the baseline values, free of interferences (claiming that the baseline represents true means without systematic biases when there is an inherent artefact). One could make the case that some deviation from the cleanness of the measure is allowed, especially in field trials, but at what point does the noise overburden the signal? In part, this is impossible to answer if no data are available to quantify the anticipatory effect. Scientific judgement then comes into play regarding the plausibility of the effect, given what the baseline population means might be. This estimation becomes subjective. What we can say, however, is that all experiments that do not consider plausible anticipatory effects, case-specific or global, are likely to produce biased before-after causal estimates to some degree.

One argument experimenters frequently make is that because members of *both* treatment and control groups were exposed to the same immediate historic and maturation effects prior to the random allocation into their respective groups, then we can conclude that these effects do not matter (Campbell et al. 1963). If there are such confounding effects, then the pretest measures represent events that the entire sample has experienced, and therefore, the between-subject tests may suffer external validity issues but not internal validity concerns. However, this assumption relies on the proposition that historic and maturation effects do not interact with the treatment effect. When there is an interaction effect, then the study introduces systematic predisposition in one arm (treatment) rather than the control arm. It will, therefore, be difficult to interpret the results of the test.

Certainly, the results of an evaluated project that has led to a reduction in crime, for example, might be desirable from a crime prevention strategy perspective (Santos 2017: p. 346). However, from a hypothesis testing view, pretest measures of crime that do not represent true unaffected-population mean scores introduce systematic bias in the causal estimates. If the study detects a significant pretest-posttest difference (or difference-in-differences), then the effect size is likely to represent an underestimated accuracy, as the observed magnitude of the differences is less than what it is in true population means. The policy implication might be that the treatment provider (the police, courts, prisons, etc.) should not invest in the tested intervention because it is more costly than other interventions. For example, the police should divest away from BWCs into other 'civilising' interventions that would reduce the use of force, complaints or arrests in police-public interactions. Put differently, BWCs may carry a larger effect size than the magnitude observed in the studies that utilise a within-subjects component in the analysis.

However, a more troubling outcome is a nonsignificant result whereby the treatment and control arms exhibit the same posttreatment response—no differences between the two arms in terms of the dependent variable. How should we interpret such results? On the one hand, the product may be proof of a true null hypothesis (e.g. 'police presence does not reduce crime in hotspots', or 'BWCs do not reduce police use of force'). On the other hand, a design-related bias is an equally plausible interpretation. This can be a result of having an underpowered study (Weisburd et al. 1993), treatment diffusion (Ariel et al. 2018) or treatment integrity concerns (see Hollin 1999: pp. 365–366). We argue in this note that another potential reason for detecting nonsignificant before-after differences is the anticipatory effect: since under both scenarios, the pretest sample conditions and posttest sample conditions, participants adjusted their behaviour *because* of the stimulus and possibly before the pretest measure was collected, then the experimenters detected nonsignificant effects. The anticipatory effects can be case-specific or global, but the result is the same, manifested in criminals freezing or displacing their criminal activity elsewhere (Perry et al. 2017); in officers reducing their proactivity prior to the pretest (Ariel et al. 2017); or by students adjusting their reported perceptions immediately prior to taking the pretest. When these conditions are met, they provide a plausible interpretation to meta-analyses that synthesise tests in which anticipatory effects contaminated either the pretest measures, created an interaction between the pretest measure and the intervention, or both, thus concluding that the treatment (e.g. police wearing BWCs) carries nil effects when such an effect does exist. In practice, the policy implications of such a review may be inaccurate.

What solutions are there for anticipatory effects?

The optimal design for assessing the accurate causal estimate is one that reduces the error rate as much as possible. Both being aware of anticipatory biases and controlling for their effects from the start of the experiment can be vital. Although it may not always be possible to quantify the anticipatory effects, especially when they are global rather than case-specific, there may be methods to minimise their influence on the study. We suggest one primary solution: collecting 'sufficiently early' pretest measures.

We first note that some experimental conditions will always remain beyond the reach of the experimenter to manipulate or to measure. Once the gorilla has been seen, it cannot be unseen, no matter what the researcher does. Instead, we propose that the means to mitigate anticipatory effects may lie in increasing the sensitivity of the analysis. If the anticipatory effect cannot be avoided, then the best option is to estimate its size. Doing so inevitably requires detailed ideas about how the mechanism(s) work and in particular an understanding of when anticipatory effects might be seen in the build-up to implementation. Experiments using official statistics such as police data, court outcomes or corrections records are likely to be in the best position to measure and then correct for anticipatory biases, by collecting multiple pretreatment observations. If the pretesting is confounded, then experimenters should collect early baseline measures that reflect the participants' behaviour prior to the appearance of the anticipatory effects. As we have shown, if the announcement of a proposed treatment caused participants to increase or reduce their behaviour in the pre-implementation period, the difference-in-difference estimator can understate the programme effect, as part of the real impact of the intervention occurred before the programme is implemented. The baseline data must, therefore, reflect the pre-announcement stage, because 'using only a small window of data around the implementation date generates biased treatment effect estimates if there are anticipation effects' (Alpert 2016: p. 32). A way to correct for this is to collect data that predates these effects.

Pushing this solution further, the 'pre-period' should cover multiple measurement points. We believe this is a generally recommended practice anyway for pretest-posttest controlled studies because it increases the explanatory power of the test over time, and it controls for potential regression to the mean, etc. However, multiple measurement points are particularly pertinent when the risk of anticipatory effects is plausible. Like retrospective longitudinal designs, the more data points collected for both the treatment and control participants, which would then show a significant 'interruption' following the treatment arm but not in the control arm, the more robust the results of the experiment will be. The repeated measures can also provide an empirical indication of the anticipatory effects.

Conclusions

We have argued that experimenters who utilise research designs with a before-after comparator must be cognisant to the possibility that the pretest measures do not provide a valid measure of the baseline. Pretest measures are assumed to be independent of the intervention, and should there be an interaction effect between the measurement(s) and the intervention, then the study results are said to be conflated: the findings do not

provide accurate causal estimates of true population means. This is the case not just for posttreatment measures—and much of the technical literature focuses on these concerns—but also for measures that are intended to estimate treatment-free, baseline conditions. We have provided examples where anticipatory effects may affect the settings of a specific study and cases where participants are affected by previous publicised trials and interventions. Scholars ought to be mindful of these effects when they plan and subsequently analyse results; one way to mitigate these biases is to collect baseline data prior to the emergence of these anticipatory effects. Although this is not always possible, we should at least be aware of the possibility that study outcomes are affected by these pretest artefacts as a method for explaining experimental results.

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., et al. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 253–267.

Allen, H. G., Stanton, T. R., Di Pietro, F., & Moseley, G. L. (2013). Social media release increases dissemination of original articles in the clinical pain sciences. *PLoS One, 8*(7), e68914.

Alpert, A. (2016). The anticipatory effects of Medicare Part D on drug utilization. *Journal of Health Economics, 49*, 28–45.

Antrobus, E., Thompson, I., & Ariel, B. (2019). Procedural justice training for police recruits: results of a randomized controlled trial. *Journal of Experimental Criminology, 15*(1), 29–53.

Ariel, B., & Partridge, H. (2017). Predictable policing: measuring the crime control benefits of hotspots policing at bus stops. *Journal of Quantitative Criminology, 33*(4), 809–833.

Ariel, B., Farrar, W. A., & Sutherland, A. (2015). The effect of police body-worn cameras on use of force and citizens' complaints against the police: a randomized controlled trial. *Journal of Quantitative Criminology, 31*(3), 509–535.

Ariel, B., Sutherland, A., Henstock, D., Young, J., Drover, P., Sykes, J., et al. (2017). "Contagious accountability" a global multisite randomized controlled trial on the effect of police body-worn cameras on citizens' complaints against the police. *Criminal Justice and Behavior, 44*(2), 293–316.

Ariel, B., Sutherland, A., & Sherman, L. W. (2018). Preventing treatment spillover contamination in criminological field experiments: the case of body-worn police cameras. *Journal of Experimental Criminology*, 1–23.

Ariel, B., Farrar, W. A., & Sutherland, A. (2019). Correction to: The effect of police body-worn cameras on use of force and citizens' complaints against the police: a randomized controlled trial. *Journal of Quantitative Criminology*, 1–2.

Barnes, J. C., TenEyck, M. F., Pratt, T. C., & Cullen, F. T. (2019). How powerful is the evidence in criminology? On Whether We Should Fear a Coming Crisis of Confidence. *Justice Quarterly*, 1–27.

Boba, R., & Roberto Santos, L. (2007). Single-family home construction site theft: a crime prevention case study. *International Journal of Construction Education and Research, 3*(3), 217–236.

Braga, A. A., Sousa, W. H., Coldren Jr., J. R., & Rodriguez, D. (2018). The effects of body-worn cameras on police activity and police-citizen encounters: a randomized controlled trial. *Journal of Criminal Law and Criminology, 108*(3), 511.

Brewer, N., & Palmer, M. A. (2010). Eyewitness identification tests. *Legal and Criminological Psychology, 15*(1), 77–96.

Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton, Mifflin and Company.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349.

Department, M. P. (2013). Program evaluation and recommendations on-officer body camera system. Retrieved from Mesa, Arizona: http://www.theiacp.org/Portals/0/documents/pdfs/LEIM/Operational%20Track%20Workshops/O2%20On%20Body%20Cameras.pdf.

Edwards, A. L. (1957). The social desirability variable in personality assessment and research.

Ellis, T., Jenkins, C., & Smith, P. (2015). Evaluation of the introduction of personal issue body worn video cameras (Operation Hyperion) on the Isle of Wight: final report to Hampshire Constabulary.

Entwisle, D. R. (1961). Interactive effects of pretesting. *Educational and Psychological Measurement, 21*(3), 607–620.

Farrington, D. P. (1983). Randomized experiments on crime and justice. *Crime and Justice, 4*, 257–308.

Farrington, D. P., Gottfredson, D. C., Sherman, L. W., & Welsh, B. C. (2003). The Maryland Scientific Methods Scale. In *Evidence-based crime prevention* (pp. 27–35). Routledge.

Farrington, D. P., Lösel, F., Boruch, R. F., Gottfredson, D. C., Mazerolle, L., Sherman, L. W., & Weisburd, D. (2018). Advancing knowledge about replication in criminology. *Journal of Experimental Criminology*, 1–24.

Franke, D., Bierie, D., & MacKenzie, D. L. (2010). Legitimacy in corrections: a randomized experiment comparing a boot camp with a prison. *Criminology & Public Policy, 9*(1), 89–117.

Gibbs, S. R. (2010). Information-generated effects. Naval Postgraduate School Monterey, CA. (thesis dissertation).

Goldkamp, J. S., & Vilcica, E. R. (2008). Targeted enforcement and adverse system side effects: the generation of fugitives in Philadelphia. *Criminology, 46*(2), 371–409.

Goodall, M. (2007). *Guidance for the police use of body-worn video devices*. London: Home Office.

Graham, A., McManus, H. D., Cullen, F. T., Burton Jr., V. S., & Jonson, C. L. (2019). Videos don't lie: African Americans' support for body-worn cameras. *Criminal Justice Review, 0734016819846229*.

Gribbons, B., & Herman, J. (1997). True and quasi-experimental designs. *Practical Assessment, Research & Evaluation, 5*(14), 1–3.

Gruijters, S. L. (2016). Baseline comparisons and covariate fishing: Bad statistical habits we should have broken yesterday. *The European Health Psychologist, 18*(5), 205–209.

Hanslmayr, S., Leipold, P., Pastötter, B., & Bäuml, K. H. (2009). Anticipatory signatures of voluntary memory suppression. *Journal of Neuroscience, 29*(9), 2742–2747.

Harvey, L. A. (2018). Statistical testing for baseline differences between randomised groups is not meaningful. *Spinal Cord, 56*(10), 919.

Haverman, R. (1987). Policy analysis and evaluation after 20 years. *Policy Studies Journal, 16*(2), 191–214.

Hohl, K., Bradford, B., & Stanko, E. A. (2010). Influencing trust and confidence in the London Metropolitan Police: results from an experiment testing the effect of leaflet drops on public opinion. *The British Journal of Criminology, 50*(3), 491–513.

Hollin, C. R. (1999). Treatment programs for offenders: meta-analysis, "what works," and beyond. *International Journal of Law and Psychiatry, 22*(3–4), 361–372.

Hollis, S., & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *British Meidcal Journal, 319*(7211), 670–674.

Jennings, W. G., Lynch, M. D., & Fridell, L. A. (2015). Evaluating the impact of police officer body-worn cameras (BWCs) on response-to-resistance and serious external complaints: evidence from the Orlando police department (OPD) experience utilizing a randomized controlled experiment. *Journal of Criminal Justice, 43*(6), 480–486.

Johnson, S. D., & Bowers, K. J. (2003). Opportunity is in the eye of the beholder: the role of publicity in crime prevention. *Criminology and public policy, 2*(3), 497–524.

Kent, R. N., O'Leary, K. D., Diament, C., & Dietz, A. (1974). Expectation biases in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology, 42*(6), 774.

Kim, E. S., & Willson, V. L. (2010). Evaluating pretest effects in pre–post studies. *Educational and Psychological Measurement, 70*(5), 744–759.

Larivière, V., Haustein, S., & Mongeon, P. (2015). The oligopoly of academic publishers in the digital era. *PLoS One, 10*(6), e0127502.

Lee, M., Taylor, E., & Willis, M. (2019). Being held to account: detainees' perceptions of police body-worn cameras. *Australian & New Zealand Journal of Criminology, 52*(2), 174–192.

Lehrer, D., Leschke, J., Lhachimi, S., Vasiliu, A., & Weiffen, B. (2007). Negative results in social science. *European Political Science, 6*(1), 51–68.

Linning, S. J., Bowers, K., & Eck, J. E. (2019). Understanding the time-course of an intervention's mechanisms: a framework for improving experiments and evaluations. *Journal of Experimental Criminology*, 1–18.

Lum, C., Stoltz, M., Koper, C. S., & Scherer, J. A. (2019). Research on body-worn cameras: what we know, what we need to know. *Criminology & Public Policy, 18*(1), 93–118.

McCambridge, J., Butor-Bhavsar, K., Witton, J., & Elbourne, D. (2011). Can research assessments themselves cause bias in behaviour change trials? A systematic review of evidence from Solomon 4-group studies. *PLoS One, 6*(10), e25223.

McNutt, M. (2014). Reproducibility. *Science (New York, NY), 343*(6168), 229.

Michelson, L., Mannarino, A., Marchione, K., Kazdin, A. E., & Costello, A. (1985). Expectancy bias in behavioural observations of therapeutic outcome: an experimental analysis of treatment and halo effects. *Behaviour Research and Therapy, 23*(4), 407–414.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., et al. (2012). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *International Journal of Surgery, 10*(1), 28–55.

Obama, B. (2016). The president's role in advancing criminal justice reform. *Harvard Law Review, 130*, 811–866.

Osca-Lluch, J., & Haba, J. (2005). Dissemination of Spanish social sciences and humanities journals. *Journal of Information Science, 31*(3), 230–237.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528–530.

Perry, S., Apel, R., Newman, G. R., & Clarke, R. V. (2017). The situational prevention of terrorism: an evaluation of the Israeli West Bank barrier. *Journal of Quantitative Criminology, 33*(4), 727–751.

Peterson, B., Yu, L., Vigne, N. L., & Lawrence, D. (2018). The Milwaukee Police Department's body-worn camera program. Retrieved from Urban Institute: https://www.urban.org/sites/default/files/publication/98461/the_milwaukee_police_departments_body_worn_camera_program_2.pdf.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., et al. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples. *British Journal of Cancer, 35*(1), 1.

President's Task Force on 21st Century Policing. (2015). *Interim report of the President's Task Force on 21st Century Policing, 28*. Washington, DC: Office of Community Oriented Policing Services Retrieved from cops.usdoj.gov/pdf/taskforce/Interim_TF_Report_150228_Intro_to_Implementation.pdf.

Pridemore, W. A., Makel, M. C., & Plucker, J. A. (2018). Replication in criminology and the social sciences. *Annual Review of Criminology, 1*, 19–38.

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15*(3), 243.

Rossi, P. H., Berk, R. A., & Lenihan, K. J. (1980). *Money, work and crime: some experimental results*. New York: Academic Press.

Santos, R. B. (2017). *Crime analysis with crime mapping*. Thousand Oaks: Sage publications.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Sheiner, L. B., & Rubin, D. B. (1995). Intention-to-treat analysis and the goals of clinical trials. *Clinical Pharmacology & Therapeutics, 57*(1), 6–15.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception, 28*(9), 1059–1074.

Smith, M. J., Clarke, R. V., & Pease, K. (2002). Anticipatory benefits in crime prevention. *Crime Prevention Studies, 13*, 71–88.

Sprangers, M., & Hoogstraten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *Journal of Applied Psychology, 74*(2), 265.

Stocking, S. H., & Dunwoody, S. L. (1982). *Social science in the mass media: images and evidence*, *In The Ethics of Social Research* (pp. 151–169). New York, NY: Springer.

Świątkowski, W., & Dompnier, B. (2017). Replicability crisis in social psychology: looking at the past to find new pathways for the future. *International Review of Social Psychology, 30*(1), 111–124.

Weisburd, D., Petrosino, A., & Mason, G. (1993). Design sensitivity in criminal justice experiments. *Crime and Justice, 17*, 337–379.

Weisburd, D., Wyckoff, L. A., Ready, J., Eck, J. E., Hinkle, J. C., & Gajewski, F. (2006). Does crime just move around the corner? A controlled study of spatial displacement and diffusion of crime control benefits. *Criminology, 44*(3), 549–592.

Weisburd, D., Farrington, D. P., & Gill, C. (Eds.). (2016). *What works in crime prevention and rehabilitation: lessons from systematic reviews*. Springer.

White, M. D., Gaub, J. E., & Todak, N. (2017). Exploring the potential for body-worn cameras to reduce violence in police–citizen encounters. *Policing: a journal of policy and practice, 12*(1), 66–76.

Yokum, D., Ravishankar, A., & Coppock, A. (2017). *Evaluating the effects of police body-worn cameras: a randomized controlled trial*. Retrieved from https://bwc.thelab.dc.gov/TheLabDC_MPD_BWC_Working_Paper_10.20.17.pdf.

Dr. Barak Ariel is an Associate Professor at Hebrew University, Jerusalem and a Lecturer in Experimental Criminology at the Institute of Criminology at Cambridge University

Dr. Alex Sutherland is Chief Scientist and Director of Research and Evaluation at the Behavioural Insight Team

Dr. Matthew Bland is a Lecturer in Evidence Based Policing at the Institute of Criminology at Cambridge University